DNA and protein sequence databanks: some caveats

by K. M. Rajkowski

It is known that conventional museums contain specimens that are mislabelled, incomplete or so damaged as to be unusable, and that some specimens will have been mislaid or lost. It is less obvious that 'silicon museums', such as Genbank and EMBL, which comprise collections of DNA and protein sequence data, also suffer from similar problems. With these problems comes the corresponding risk of compromising increasingly popular molecular phylogenetic studies. Here, types of problem and reasons for suspicion of certain sequences, as well as examples of the inutility of others, are outlined and methods for detecting some errors proposed.

Examples were found in the 2,470 mitochondrial cytochrome b sequence entries obtained by searching the EMBL v.60.0/EMBLNEW v.61.0 databank, updated to 3 November 1999 (5,087,527 sequences), using the keyword 'cytochrome b'. Each entry consists of a *label*, identifying and describing the origins of the sequence, and a *specimen*, the DNA and translated protein sequences.

Problem entries included: (1) specimens too fragmentary for use—as few as 13 nucleotide base pairs. Only a minority of specimens are of the minimum number of base pairs (approx. 900) needed for protein molecular phylogenies when other genes for the correponding species have not been sequenced (as is mostly the case); (2) specimens where the proportion of undetermined nucleotides was so large that the sequence was unusable; (3) specimens where the DNA sequence had been translated into the protein sequence using the wrong genetic code (nuclear instead of mitochondrial); and (4) specimens not identified to species level (i.e. unlabelled). Mislaid specimens, not found in the search but present in the databank, included 14 *Phylloscopus* (leaf-warbler) sequences (because of a typing error in their labels ['cytochorome b']), and 44 sequences for Corvidae (crows), Sylviidae (warblers) and Timaliidae (babblers) only found using other keywords. Lost specimens included seven *Phylloscopus* sequences published in 1992 but still absent from the databanks. Furthermore, many entries were confirmatory replicates that could be combined into a single sequence entry with the corresponding annotations on the label.

While many such problem entries should be detectable with appropriate computer programmes rather than 'manually' searching through (in the case of cytochrome b sequences) some 3,000 pages of text, the problem of sequencing errors remains. It is estimated that, on average, 0.1% of the nucleotides in databanks are mis-sequenced and, for methodological reasons, errors will be more frequent in some sequences than others. Some probable nucleotide sequencing errors are detectable provided they give rise to improbable amino acid substitutions in the translated protein, and for this reason it is recommended that the protein sequence be compared with those of related species prior to using the nucleotide sequence in a phylogenetic study or, preferably, before submitting it to a databank.

Address : K. M. Rajkowski, INSERM Unité 488, Hôpital de Bicêtre, F-94276 Le Kremlin-Bicêtre, France.